What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in Web argumentation

Ivan Habernal† and Iryna Gurevych†‡
†Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science, Technische Universität Darmstadt
‡Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract
This article tackles a new challenging task in computational argumentation. Given a pair of two arguments to a certain controversial topic, we aim to directly assess qualitative properties of the arguments in order to explain why one argument is more convincing than the other one. We approach this task in a fully empirical manner by annotating 26k explanations written in natural language. These explanations describe convincingness of arguments in the given argument pair, such as their strengths or flaws. We create a new crowd-sourced corpus containing 9,111 argument pairs, multi-labeled with 17 classes, which was cleaned and curated by employing several strict quality measures. We propose two tasks on this data set, namely (1) predicting the full label distribution and (2) classifying types of flaws in less convincing arguments. Our experiments with feature-rich SVM learners and Bidirectional LSTM neural networks with convolution and attention mechanism reveal that such a novel fine-grained analysis of Web argument convincingness is a very challenging task. We release the new corpus UKPConvArg2 and the accompanying software under permissive licenses to the research community.

1 Introduction
People engage in argumentation in various contexts, both online and in the real life. Existing definitions of argumentation do not solely focus on giving reasons and laying out a logical framework of premises and conclusions, but also highlight its social purpose which is to convince or to persuade (O’Keefe, 2011; van Eemeren et al., 2014; Blair, 2011). Assessing the quality and strength of perceived arguments therefore plays an inherent role in argumentative discourse. Despite strong theoretical foundations and plethora of normative theories, such as Walton’s schemes and their critical questions (Walton, 1989), an ideal model of critical discussion in the pragma-dialectic view (Van Eemeren and Groothendijk, 1987), or research into fallacies (Boudry et al., 2015), assessing qualitative criteria of everyday argumentation represents a challenge for argumentation scholars and practitioners (Weltzer-Ward et al., 2009; Swanson et al., 2015; Rosenfeld and Kraus, 2015).

Addressing qualitative aspects of arguments has recently started gaining attention in the field of computational argumentation. Scoring strength of persuasive essays (Farra et al., 2015; Persing and Ng, 2015), exploring interaction in persuasive dialogues on Reddit (Tan et al., 2016), or detecting convincing arguments (Habernal and Gurevych, 2016) are among recent attempts to tackle the quality of argumentation. However, these approaches are holistic and do not necessarily explain why a given argument is strong or convincing.

We asked the following research questions. First, can we assess what makes an argument convincing in a purely empirical fashion as opposite to theoretical normative approaches? Second, to what extent can the problem be tackled by computational models? To address these questions, we exploit our recently introduced UKPConvArg1 corpus (Habernal and Gurevych, 2016). This data set consists of 11,650 argument pairs – two arguments with the
**Prompt:** Should physical education be mandatory in schools? **Stance:** Yes!

<table>
<thead>
<tr>
<th>Argument 1</th>
<th>Argument 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE should be compulsory because it keeps us constantly fit and healthy. If you really dislike sports, then you can quit it when you’re an adult. But when you’re a kid, the best thing for you to do is study, play and exercise. If you prefer to be lazy and lie on the couch all day then you are most likely to get sick and unfit. Besides, PE helps kids be better at teamwork.</td>
<td>physical education should be mandatory cuhz 112,000 people have died in the year 2011 so far and it’s because of the lack of physical activity and people are becoming obese!!!!</td>
</tr>
</tbody>
</table>

A1 is more convincing than A2, because:
- “A1 is more intelligently written and makes some good points (teamwork, for example). A2 used ‘cuhz’ and I was done reading because that sounds stupid.”
- “A1 gives more reasons and goes into detail, A2 only has one fact”
- “A1 makes several compelling points. A2 uses poor spelling and grammar.”

**Figure 1:** An annotated argument pair from the UKPConvArg corpus with three reasons explaining the decision about convincingness (ID arg54258_arg202285).

We use these natural language reasons as a proxy to assess qualitative properties of the arguments in each argument pair. Our main contributions are: (1) We propose empirically inspired labels of quality properties of Web arguments and design a hierarchical annotation scheme. (2) We create a new large crowd-sourced benchmark data set containing 9,111 argument pairs multi-labeled with 17 categories which is improved by local and global filtering techniques. (3) We experiment with several computational models, both traditional and neural network-based, and evaluate their performance quantitatively and qualitatively.

The newly created data set UKPConvArg2 is available under CC-BY-SA license along with the experimental software for full reproducibility at GitHub.1

2 Related Work

The growing field of computational argumentation has been traditionally devoted to structural tasks, such as argument component detection and classification (Habernal and Gurevych, 2017; Habernal and Gurevych, 2015), argument structure parsing (Peldszus and Stede, 2015; Stab and Gurevych, 2014), or argument schema classification (Lawrence and Reed, 2015), leaving the issues of argument evaluation or quality assessment as an open future work.

There are only few attempts to tackle the qualitative aspects of arguments, especially in the Web discourse. Park and Cardie (2014) classified propositions in Web arguments into four classes with respect to their level of verifiability. Focusing on convincingness of Web arguments, Habernal and Gurevych (2016) annotated 16k pairs of arguments with a binary relation “is more convincing” and also elicited explanation for the annotators’ decisions.

Recently, research in persuasive essay scoring has started combining holistic approaches based on rubrics for several dimensions typical to this genre with explicit argument detection. Persing and Ng (2015) manually labeled 1,000 student persuasive essays with a single score on the 1–4 scale and trained a regression predictor with a rich feature set using LIBSVM. Among traditional features (such as POS or semantic frames), an argument structure parser by Stab and Gurevych (2014) was employed. Farra et al. (2015) also deal with essay scoring but rather then tackling the argument structure, they focus on methods for detecting opinion expressions. Persuasive essays however represent a genre with a rather strict qualitative and formal requirements (as taught in curricula) and substantially differ from online argumentation.

Argument evaluation belongs to the central research topics among argumentation scholars (Toul-
Yet treatment of assessing argumentation quality, persuasiveness, or convincingness is traditionally based on evaluating relevance, sufficiency or acceptability of premises (Govier, 2010; Johnson and Blair, 2006) or categorizing fallacies (Hamblin, 1970; Tindale, 2007). However, the nature of these normative approaches causes a gap between the ‘ideal’ models and empirically encountered real-world arguments, such as those on the Web (van Eemeren et al., 2014; Walton, 2012).

Regarding the methodology utilized later in this paper, deep (recursive) neural networks have gained extreme popularity in NLP in recent years. Long Short-Term Memory networks (LSTM) with Attention mechanism have been applied on textual entailment (Rocktäschel et al., 2016), Question-Answering (Golub and He, 2016), or source-code summarization (Allamanis et al., 2016).

3 Data

As our source data set, we took the publicly available UKPConvArg1 corpus. It is based on arguments originated from 16 debates from Web debate platforms createdebate.com and convinceme.net, each debate has two sides (usually pro and con). Arguments from each of the 32 debate sides are connected into a set of argument pairs, and each argument pair is annotated with a binary relation (argument A is more/less convincing than argument B), resulting in total into 11,650 argument pairs. Annotations performed by Habernal and Gurevych (2016) also contain several reasons written by crowd-workers that explain why a particular argument is more or less convincing; see an example in Figure 1.

As these reasons were written in an uncontrolled setting, they naturally reflect the main properties of argument quality in a downstream task, which is to decide which argument from a pair is more convincing. It differs from scoring arguments in isolation, which is inherently harder not only due to subjectivity in argument “strength” decision but also because of possible annotator’s prior bias (Habernal and Gurevych, 2016). Assessing an argument in context helps to emphasize its main flaws or strengths. This approach is also known as knowledge elicitation – acquiring appropriate information from experts by asking "why?” (Reed and Rowe, 2004).

We therefore used the reasons as a proxy for developing a scheme for labeling argument quality attributes. This was done in a purely bottom-up empirical manner, as opposed to using ‘standard’ evaluation criteria known from argumentation literature (Johnson and Blair, 2006; Schiappa and Nordin, 2013). In particular, we split all reasons into several reason units by simple preprocessing (splitting using Stanford CoreNLP (Manning et al., 2014), segmentation into Elementary Discourse Units by RST tools (Surdeanu et al., 2015)) and identified the referenced arguments (A1 or A2) by pattern matching and dependency parsing. For example, each reason from Figure 1 would be transformed into two reason units. Overall, we obtained about 70k reason units from the entire UKPConvArg1 corpus.

3.1 Annotation scheme

In order to develop a code book for assigning a label to each reason unit, we ran several pilot expert annotation studies (each with 200-300 reason units). Having a set of ≈ 25 distinct labels, we ran two larger studies on Amazon Mechanical Turk (AMT), each with 500 reason units and 10 workers. The workers were split into two groups; we then estimated gold labels for each group using MACE (Hovy et al., 2013) and compared both groups’ results in order to find systematic discrepancies. Finally, we ended up with a set of 19 distinct labels (classes). As the number of classes is too big for non-expert crowd workers, we developed a hierarchical annotation process guided by questions that narrow down the final class decision. The scheme is depicted in Figure 2. Workers were shown only the reason units without seeing the original arguments.

3We picked this example for its simplicity, in reality the texts are much more fuzzy.

4It might seem that some labels are missing, such as C8-2 and C8-3; these belong to those removed during the pilot studies.
3.2 Annotation

We sampled 26,000 unique reason units ordered by the original author competence provided as part of the UKPComArg corpus. We expected that workers with higher competence tend to write better reasons for their explanations. Using the previously introduced scheme, 776 AMT workers annotated the batch during two weeks; we required assignments from 5 workers for a single item. We employed MACE (Hovy et al., 2013) for gold label and worker competence estimation with 95% threshold to ignore the less confident labels. Several workers were rejected based on their low computed competence and other criteria, such as too short submission times.

3.3 Data cleaning

We performed several cleaning procedures to increase quality and consistency of the annotated data (apart from initial MACE filtering already explained above).

Local cleaning First, we removed 3,859 reason units annotated either with C1-2 (“not an explanation”) and C8-6 (“too topic-specific”, which usually paraphrases some details from the related argument and is not general enough). In the next step, we removed reason units with wrong polarity. In particular, all reason units labeled with C8-* or C9-* should refer to the more convincing argument in the argument pair (as they describe positive properties), whereas all reasons with labels C5-*, C6-*, and C7-* should refer to the less convincing argument. The target arguments for reason units were known from the heuristic preprocessing (see above); in this step 2,455 units were removed.

Global cleaning Since the argument pairs from one debate can be projected into an argument graph (Habernal and Gurevych, 2016), we utilized this ‘global’ context for further consistency cleaning.

Suppose we have two argument pairs, $P_1(A \rightarrow B)$ and $P_2(B \rightarrow C)$ (where $\rightarrow$ means “is more convincing than”). Let $P_1(R_B)$ be reason unit targeting
\( B \) in argument pair \( P_1 \) and similarly \( P_2(R_B) \) reason unit targeting \( B \) in argument pair \( P_2 \). In other words, two reason units target the same argument in two different argument pairs (in one of them the argument is more convincing while in the other pair it is less convincing). There might then exist contradicting combination of classes for \( P_1(R_B) \) and \( P_2(R_B) \). For example classes C9-2 and C7-3 are contradicting, as the same argument cannot be both "on the topic" and "off-topic" at the same time.

When such a conflict between two reason units occurred, we selected the reason with a higher score using the following formula:

\[
w_W \ast \sigma \left( \sum_{A=G} w_A - \lambda \sum_{A \neq G} w_A \right) \tag{1}
\]

where \( w_W \) is the competence of the original author of the reason unit (originated from the UKP-ConvArg corpus), \( A = G \) are crowdsourced assignments for a single reason unit that match the final predicted gold label, \( A \neq G \) are assignments that differ from the final predicted gold label, \( w_A \) is the competence of worker for assignment \( A \), \( \lambda \) is a penalty for non-gold labels, and \( \sigma \) is the sigmoid function to squeeze the score between 0 and 1.

We found 25 types of global contradictions between labels for reason units and used them for cleaning the data; in total 3,790 reason units were removed in this step. After all cleaning procedures, annotations from reason units were mapped back to argument pairs, resulting into a multi-label annotation of one or both arguments from the given pair. In total 9,111 pairs from the UKPConvArg corpus were annotated.

For example, the final annotations of argument pair shown in Figure 1 contain four labels – C8-1 (as the more convincing argument "has more details, information, facts, or examples / more reasons / better reasoning / goes deeper / is more specific"), C9-3 (as the more convincing argument "has provoking question / makes you think"), C5-2 (as the less convincing argument "has language issues / bad grammar /..."), and C6-1 (as the less convincing argument "provides not enough support / ..." ). Only four of six reason units for this argument pair were annotated because of the competence score of their authors.

Table 1 shows number of labels per argument pairs; about a half of the argument pairs have only one label. Figure 3 shows distribution of label in the entire data set which is heavily skewed towards C8-1 label. This is not surprising, as this label was used for reason units pointing out that the more convincing argument provided more reasons, details, information or better reasoning – a feature inherent to argumentation seen as giving reasons (Freeley and Steinberg, 2008).

### 3.4 Data validation

Since the qualitative attributes of arguments were annotated indirectly by labeling their corresponding reason units without seeing the original arguments, we wanted to validate correctness of this approach. We designed a validation study, in which workers were shown the original argument pair and two sets of labels. The first set contained the true labels as annotated previously, while we randomly replaced few labels in the second set. The goal was then to decide which set of labels better explains that argument A is...
more convincing than argument B. For example, for the argument pair from Figure 1, one set of shown labels would be \{C8-1, C9-3, C5-2, C6-1\} (the correct set) while the other ‘distracting’ set would be \{C8-1, C9-3, C5-1, C7-3\}.

We randomly sampled 500 argument pairs and collected 9 assignments per pair on AMT; we again used MACE with 95% threshold. Accuracy of workers on 235 argument pairs achieved 82%. We can thus conclude that workers tend to prefer explanations based on labels from the reason units and using the annotation process presented in this section is reliable. Total costs of the annotations including pilot studies, bonuses, and data validation were USD 3,300.

4 Experiments

We propose two experiments, both performed in 16-fold cross-domain validation. In each fold, argument pairs from 15 debates are used and the remaining one is used for testing. In both experiments, it is assumed that the more convincing argument in a pair is known and we concatenate (using a particular delimiter) both arguments such that the more convincing argument comes first.

4.1 Predicting full multi-label distribution

This experiment is a multi-label classification. Given an argument pair annotated with several labels, the goal is to predict all these labels.

We use two deep learning models. Our first model, Bidirectional Long Short-Term Memory (BLSTM) network contains two LSTM blocks (forward and backward), each with 64 hidden units on the output. The output is concatenated into a single vector and pushed through sigmoid layer with 17 output units (corresponding to 17 labels). We use cross entropy loss function in order to minimize distance of label distributions in training and test data (Nam et al., 2014). In the input layer, we rely on pre-trained word embeddings from Glove (Pennington et al., 2014) whose weights are updated during training the network.

The second models is BLSTM extended with an attention mechanism (Rocktäschel et al., 2016; Golub and He, 2016) combined with convolution layers over the input. In particular, the input embedding layer is convoluted using 4 different convolution sizes (2, 3, 5, 7), each with 1,000 randomly initialized weight vectors. Then we perform max-over-time pooling and concatenate the output into a single vector. This vector is used as the attention module in BLSTM.

We evaluate the system using two widely used metrics in multi-label classification. First, Hamming-loss is the average per-item per-class total error; the smaller the better (Zhang and Zhou, 2007). Second, we report One-error (Sokolova and Lapalme, 2009) which corresponds to the error of the predicted label with highest probability; the smaller the better.

We do not report other metrics (such as Area Under PRC-curves, MAP, or cover) as they require tuning a threshold parameter, see a survey by Zhang and Zhou (2014).

Results from Table 2 do not show significant differences between the two models. Putting the one-error numbers into human performance context can be done only indirectly, as the data validation pre-

<table>
<thead>
<tr>
<th>Debate</th>
<th>BLSTM H-loss</th>
<th>one-E</th>
<th>BLSTM/CNN/ATT H-loss</th>
<th>one-E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ban plastic water bottles?</td>
<td>0.092</td>
<td>0.283</td>
<td>0.090</td>
<td>0.305</td>
</tr>
<tr>
<td>Christianity or Atheism</td>
<td>0.105</td>
<td>0.212</td>
<td>0.105</td>
<td>0.218</td>
</tr>
<tr>
<td>Evolution vs. Creation</td>
<td>0.093</td>
<td>0.196</td>
<td>0.094</td>
<td>0.234</td>
</tr>
<tr>
<td>Firefox vs. Internet Explorer</td>
<td>0.080</td>
<td>0.312</td>
<td>0.078</td>
<td>0.345</td>
</tr>
<tr>
<td>Gay marriage: right or wrong?</td>
<td>0.095</td>
<td>0.243</td>
<td>0.094</td>
<td>0.270</td>
</tr>
<tr>
<td>Should parents use spanking?</td>
<td>0.082</td>
<td>0.312</td>
<td>0.083</td>
<td>0.344</td>
</tr>
<tr>
<td>If your spouse committed murder...</td>
<td>0.094</td>
<td>0.297</td>
<td>0.094</td>
<td>0.272</td>
</tr>
<tr>
<td>India has the potential to lead the world</td>
<td>0.088</td>
<td>0.294</td>
<td>0.086</td>
<td>0.322</td>
</tr>
<tr>
<td>Is it better to have a lousy father or to be fatherless?</td>
<td>0.086</td>
<td>0.367</td>
<td>0.085</td>
<td>0.381</td>
</tr>
<tr>
<td>Is porn wrong?</td>
<td>0.098</td>
<td>0.278</td>
<td>0.100</td>
<td>0.270</td>
</tr>
<tr>
<td>Is the school uniform a good or bad idea?</td>
<td>0.081</td>
<td>0.279</td>
<td>0.077</td>
<td>0.406</td>
</tr>
<tr>
<td>Pro-choice vs. Pro-life</td>
<td>0.095</td>
<td>0.218</td>
<td>0.098</td>
<td>0.218</td>
</tr>
<tr>
<td>Should Physical Education be mandatory?</td>
<td>0.095</td>
<td>0.273</td>
<td>0.095</td>
<td>0.277</td>
</tr>
<tr>
<td>TV is better than books</td>
<td>0.091</td>
<td>0.265</td>
<td>0.087</td>
<td>0.300</td>
</tr>
<tr>
<td>Personal pursuit or common good?</td>
<td>0.095</td>
<td>0.328</td>
<td>0.094</td>
<td>0.343</td>
</tr>
<tr>
<td>W. Farquhar ought to be honored...</td>
<td>0.054</td>
<td>0.528</td>
<td>0.052</td>
<td>0.570</td>
</tr>
<tr>
<td>Average</td>
<td>0.089</td>
<td>0.293</td>
<td>0.088</td>
<td>0.317</td>
</tr>
</tbody>
</table>

Table 2: Results of multi-label classification from Experiment 1. Hamming-loss and One-Error are shown for two systems – Bidirectional LSTM and Bidirectional LSTM with Convolution and Attention.
Error analysis and discussion  We examined outputs from the label distribution prediction for BLSTM/ATT/CNN. It turns out that the output layer leans toward predicting the dominant label C8-1, while prediction of other labels is seldom. We suspect two causes, first, the highly skewed distribution of labels (see Figure 3) and, second, insufficient training data sizes where 13 classes have less than 1k training examples (while Goodfellow et al. (2016) recommend at least 5k instances per class).

Although multi-label classification may be viewed as a set of binary classification tasks that decides for each label independently (and thus allows for employing other ‘standard’ classifiers such as SVM), this so-called binary relevance approach ignores dependencies between the labels. That is why we focused directly on deep-learning methods, as they are capable of learning and predicting a full label distribution (Nam et al., 2014).

4.2 Predicting flaws in less convincing arguments

In the second experiment, we focus on predicting flaws in arguments using coarse-grained labels. While this task makes several simplifications in the labeling, it still provides meaningful insights into argument quality assessment. For this purpose, we use only argument pairs where the less convincing argument is labeled with a single label (no multi-label classification). Second, we merged all labels from categories C5-* C6-* C7-* into three classes corresponding to their parent nodes in the annotation decision schema from Figure 2. Table 3 shows distribution of the gold data for this task with explanation of the labels. It is worth noting that predicting flaws in the less convincing argument is still context-dependent and requires the entire argument pair because some of the quality labels are relative to the more convincing argument (such as “less reasoning” or “not enough support”).

<table>
<thead>
<tr>
<th>Label</th>
<th>Instances</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5</td>
<td>856</td>
<td>Language and presentation issues</td>
</tr>
<tr>
<td>C6</td>
<td>1,203</td>
<td>Reasoning and factuality issues</td>
</tr>
<tr>
<td>C7</td>
<td>1,651</td>
<td>Off-topic, non-argument, nonsense</td>
</tr>
<tr>
<td>Total</td>
<td>3,710</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Gold data distribution for the second experiment. Argument pairs with a single label for the less convincing argument.

The non-linear output function is softmax and we train the networks using categorical cross-entropy loss. We also add another baseline model that employs SVM with RBF kernel and a rich set of linguistically motivated features, similarly to (Habernal and Gurevych, 2016). The feature set includes uni- and bi-gram presence, ratio of adjective and adverb endings that may signalize neuroticism (Corney et al., 2002), contextuality measure (Heylighen and Dewaele, 2002), dependency tree depth, ratio of exclamation or quotation marks, ratio of modal verbs, counts of several named entity types, ratio of past vs. future tense verbs, POS n-grams, presence of dependency tree production rules, seven different readability measures (e.g., Ari (Senter and Smith, 1967), Coleman-Liau (Coleman and Liau, 1975), Flesch (Flesch, 1948), and others), five sentiment scores (from very negative to very positive) (Socher et al., 2013), spell-checking using standard Unix words, ratio of superlatives, and some surface features such as sentence lengths, longer words count, etc. It results into a sparse 60k-dimensional feature vector space.

Results in Table 4 suggest that the SVM-RBF baseline system performs poorly and its results are on par with a majority class baseline (not reported in detail). Both deep learning models significantly outperform the baseline, yielding Macro-$F_1$ score about 0.35. The attention-based model performs better than simple BLSTM in two classes (C5 and C6), but the overall Macro-$F_1$ score is not significantly better.

5We used LISBVM (Chang and Lin, 2011) with the default hyper-parameters. As Fernández-Delgado et al. (2014) show, SVM with gaussian kernels is a reasonable best choice on average.

6Detailed explanation of the features can be found directly in the attached source codes.
### Table 4: Results for experiment 2. P = precision, R = recall, M-\(F_1\) = macro \(F_1\), C.I. = confidence interval at 0.95. Both BLSTM and BLSTM/ATT/CNN are significantly better than SVM-RBF (\(p < 0.05\), exact Liddell’s test).

<table>
<thead>
<tr>
<th>Model</th>
<th>Class C5</th>
<th></th>
<th></th>
<th></th>
<th>Class C6</th>
<th></th>
<th></th>
<th></th>
<th>Class C7</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>(F_1)</td>
<td>P</td>
<td>R</td>
<td>(F_1)</td>
<td>P</td>
<td>R</td>
<td>(F_1)</td>
<td>M-(F_1)</td>
<td>C.I.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM-RBF</td>
<td>0.351</td>
<td>0.023</td>
<td>0.044</td>
<td>0.394</td>
<td>0.083</td>
<td>0.137</td>
<td>0.446</td>
<td>0.918</td>
<td>0.600</td>
<td>0.260</td>
<td>0.014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLSTM</td>
<td>0.265</td>
<td>0.600</td>
<td>0.368</td>
<td>0.376</td>
<td>0.229</td>
<td>0.285</td>
<td>0.479</td>
<td>0.301</td>
<td>0.370</td>
<td>0.341</td>
<td>0.015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLSTM/ATT/CNN</td>
<td>0.270</td>
<td>0.625</td>
<td>0.378</td>
<td>0.421</td>
<td>0.247</td>
<td>0.311</td>
<td>0.484</td>
<td>0.301</td>
<td>0.371</td>
<td>0.353</td>
<td>0.015</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Error analysis** We manually examined several dozens of predictions where the BLSTM model failed but the BLSTM/ATT/CNN model was correct in order to reveal some phenomena that the system is capable to cope with. First, the BLSTM/ATT/CNN model started catching some purely abusive, sarcastic, and attacking arguments. Also, the language/grammar issues were revealed in many cases, as well as using slang in arguments.

Examining predictions in which both systems failed reveal some fundamental limitations of the current purely data-driven computational approach. While the problem of not catching off-topic arguments can be probably modeled by incorporating the debate description or some sort of debate topic model into the attention vector, the more common issue of non-sense arguments or fallacious arguments (which seem like actual arguments on the first view) needs much deeper understanding of real-world knowledge, logic, and reasoning.

### 5 Conclusion

This paper presented a novel task in the field of computational argumentation, namely empirical assessment of reasons for argument convincingness. We created a new large benchmark data set by utilizing a new annotation scheme and several filtering strategies for crowdsourced data. Then we tackled two challenging tasks, namely multi-label classification of argument pairs in order to reveal qualitative properties of the arguments, and predicting flaws in the less convincing argument from the given argument pair. We performed all evaluations in a cross-domain scenario and experimented with feature-rich SVM and two state-of-the-art neural network models. The results are promising but show that the task is inherently complex as it requires deep reasoning about the presented arguments that goes beyond capabilities of the current computational models. By releasing the UKPConVArg2 data and code to the community, we believe more progress can be made in this direction in the near future.

**Acknowledgments**

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant N°2/182806, by the German Institute for Educational Research (DIPF), by the German Research Foundation (DFG) via the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1), by the GRK 1994/1 AIPHES (DFG), by the ArguAna Project GU 798/20-1 (DFG), and by Amazon Web Services in Education Grant award. Lastly, we would like to thank the anonymous reviewers for their valuable feedback.

**References**


Malcolm Corney, Olivier de Vel, Alison Anderson, and George Mohay. 2002. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th An*


